NetSPI™

AI/ML Penetration Testing

LLM model pentesting to reduce the risk of using AI in your environment

The most trusted products, services, and brands are secured by NetSPI

The Challenge

52% of data security leaders are concerned about the possibility of AI attacks via threat actors, and 57% report an increase in AI-driven attacks in the last year.[1]

According to McKinsey's latest Global Survey on AI[2], 65% of respondents regularly use AI, almost double the number of respondents from the previous year. However, although companies are eager to use AI, not every company understands the associated risks. Whether you are fine tuning off-the-shelf models, using large language learning model functionality in your applications, or in other processes, security should not be an afterthought.

The ability to identify vulnerabilities specific to LLM capabilities is critical, especially when incorporating AI into application development. Security and privacy are significant concerns. Lack of proper evaluation may allow users to manipulate LLMs, such as chatbots and expose sensitive data, generate unauthorised content, or take actions on their behalf.

The Solution

NetSPI AI/ML Penetration Testing solves these challenges using a powerful combination of people, processes, and technology, and helps reduce the risk of using AI in your environment. NetSPI offers a depth and breadth of testing, whether you need to securely incorporate LLM capabilities into your web-facing applications, gain detailed benchmarking and analysis of potential jailbreak consequences of your LLM, or customise an advanced model evaluation and review. Our rigorous and consistent testing methodology ensures we find vulnerabilities, exposures, and misconfigurations that others miss.

- Pentest LLM web applications
- Benchmark and jailbreak testing for LLMs
- Customised testing for LLM deep model evaluation

96% of executives say adopting generative AI makes a security breach likely in their organisation within the next three years.[3]

1. Immuta: The AI Security & Governance Report, Published: April 29, 2024.
2. McKinsey: The state of AI in early 2024, Published: May 30, 2024.
3. IBM: Institute for Business Value, Published: June 14, 2024.

NetSPI™

Shift Left and Uncover Web Application LLM Security Vulnerabilities Prior to Production

Without proper system configuration and security measures, LLM capabilities, such as those found in chatbots, can be exploited for malicious actions or to leak private information. Continuous testing ensures that as your application development and models evolve, you can stay ahead in identifying and mitigating vulnerabilities.

- Save time and resources by identifying exploits during development
- Uncover risks to LLM capabilities not found by static and dynamic testing
- Depth and breadth of testing for LLMs in any framework

Gain Benchmarking and Analysis of Potential Jailbreak Consequences of Your LLM

Security and privacy have become significant concerns as more applications and SaaS providers adopt LLM capabilities. These new features may allow users to manipulate LLMs, such as chatbots and expose sensitive data, generate unauthorised content, or take actions on their behalf.

- Assess and enhance your resilience against real-world threats to your LLM
- Evaluate your LLM with monthly testing, including security metrics and trend data
- Expand beyond traditional security and understand risk of LLM manipulation

Enable a Deep Advanced Model Evaluation and Review of Your LLM

Predictive and custom models within applications need deeper analysis. NetSPI can deliver a deep review of the data collection, training data structure and cleaning, training data validation, and algorithms of your model. Evaluation can also be performed to test, including but not limited to, advanced model extraction, member attribution, inference, inversion, and evasion attacks.

- Understand the impacts of usability, bias, and fairness of your LLM
- Gain deeper understanding of model weakness and controls for mitigation
- Improve the overall security of your LLM

About NetSPI

NetSPI® pioneered Penetration Testing as a Service (PTaaS) and leads the industry in modern pentesting. Combining world-class security professionals with AI and automation, NetSPI delivers clarity, speed, and scale across 50+ pentest types, attack surface management, and vulnerability prioritisation. The NetSPI platform streamlines workflows and accelerates remediation, enabling our experts to focus on deep dive testing that uncovers vulnerabilities others miss. Trusted by the top 10 U.S. banks and Fortune 500 companies worldwide, NetSPI has been driving security innovation since 2001. NetSPI is headquartered in Minneapolis, MN, and available on **AWS Marketplace**. Follow us on **LinkedIn** and **X**.